Performance Tuning of a High Capacity/High Performance Archive for the Earth Observing Systems Project

Alla Lake
Lockheed Martin Space Mission Systems and Services
1616 McCormick Drive
Upper Marlboro, MD 20774
alake@eos.hitc.com
tel: +1-301-925-0626
fax: +1-301-925-0651

Abstract: The Archive for the Earth Observing Systems (EOS) Project reflects the scale of the project itself. Its data holdings over all the Distributed Active Archive Centers (DAACs) are projected to exceed two Petabytes by the year 2002. This paper describes the experience and results of the integration and tuning process for the hardware and Commercial Off The Shelf (COTS) software used today in the EOS Core System (ECS)[1] archive. Sizable performance improvements have been realized to date through the combined efforts of the vendors and the tuning team. The process is still continuing and, indeed, will continue through the life of the archive. The effort to date centered around improving the throughput for the individual data streams to the archive tape drives and the cumulative throughput to all the drives. The work was specifically directed towards the COTS hardware and software integration and tuning adjustments. Throughput performance improvement via efficient data organization on tape was not addressed during this effort, but will be in the future work.

1 Introduction

ECS archive hardware is currently installed at five distributed sites, known as Distributed Active Archive Centers, or DAACs: Goddard Space Flight Center (GSFC) in Greenbelt, Maryland, Langley Research Center (LaRC) in Hampton, Virginia, Earth Resources Observation System (EROS) Data Center (EDC) in Sioux Falls, South Dakota, Jet Propulsion Laboratory (JPL) in Pasadena, California, and the National Snow and Ice Data Center (NSIDC) in Boulder, Colorado.

Among these sites, GSFC is the largest in both the accumulation of data and in the daily data throughput. The data holdings at GSFC alone will amount to more than one Petabyte in 2002. The data rates through the DAAC archives will match the storage capacity scale of the archives at the corresponding sites. For example, per current baseline, at its peak data ingestion period beginning in the year 2000, GSFC DAAC is to ingest close to a Terabyte of data per twenty four-hour period. This includes all levels of products, from L0 through the progression of higher level products for permanent storage and the data produced by reprocessing. To satisfy output requirements, the system is designed to distribute data to users at approximately twice the ingestion rate.

Not surprisingly, the capability of the archive to absorb and to serve data at these rates depends on a well-integrated, well-tuned combination of high performance hardware and software. The task of building an archive for ECS is complicated by the amount of data accumulating in permanent storage. In order to accommodate the expected volume of data, the storage must provide high capacity, in addition to high throughput - a difficult combination. Another complicating factor is the necessity of building the archive with components that were commercially available at the time of selection and procurement, i.e. 1995, 1996.

All the ECS DAACs have the same architecture and constituent components. The DAACs differ only in the size and particulars of equipment. Therefore, properly integrating and tuning the largest site, GSFC, creates a configuration template for most of the remaining DAACs.

Figure 1, ECS Architecture, illustrates the relative position of the archive within ECS configuration and the data flows within a DAAC.

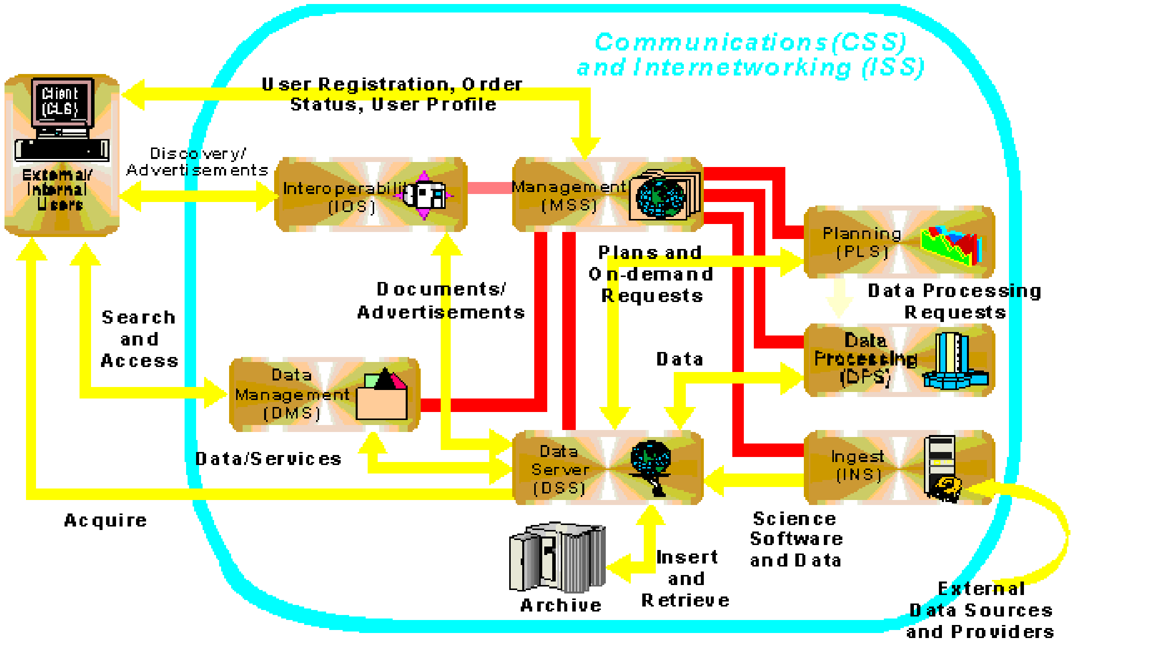


Figure 1. ECS Architecture

This paper describes integration and tuning efforts for the archive repository component of the GSFC DAAC, labeled **Archive** in Figure 1.

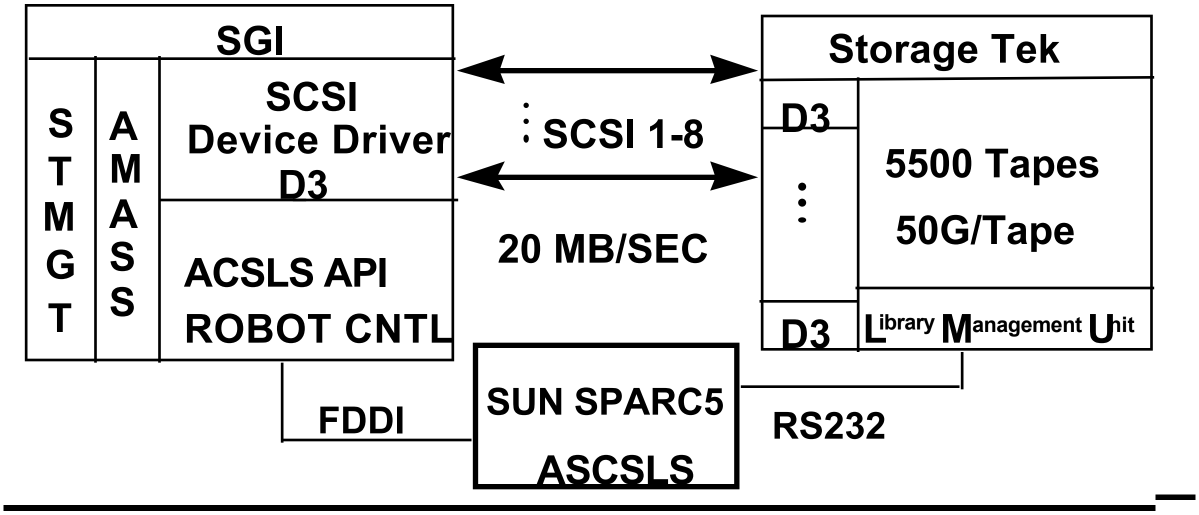
The integration and performance tuning that were undertaken could be described as a series of movements of the data from one component of the archive to another: from the data rate achieved on an individual tape drive, to the exchange rate of the robotic mechanism, to the overall data rate to and from the entire silo. The shifts were largely realized by improvements in the software design and function and by gaining additional performance from the RAID. Some of the desired enhancements are still pending at the time of this writing in November of 1997.

2 Architecture and Equipment Suite at the GSFC DAAC Archive Repository

The integration and tuning work described below pertains to the following suite of commercial products. AMASS[2] File Storage Management System (FSMS) software, from EMASS Corporation, controls the physical storage of the data collection and is hosted on a multiprocessor Silicon Graphics Challenge server. The data repository resides in STK PowderHorn robotic silos and is recorded using RedWood (SD-3) helical scan tape drives from STK Corporation. Redundant Array of Inexpensive Disks (RAID) from Silicon Graphics Corporation, configured as RAID-3, is used for the temporary caching of data en route to and from the robotic silos. The initial (at-launch) GSFC archive configuration consists of two STK PowderHorn robotic silos. Each silo is equipped with eight SD-3 helical scan tape drives.

The tape drives residing in the STK robotic silo are directly connected to the SGI host via Fast-And-Wide SCSI II channels, one channel per drive. Each channel is individually capable of the throughput of 20 MB/sec. Each of the eight tape drives is rated by the manufacturer as capable of 11.2 MB/sec sustained throughput. The drives exhibited even higher streaming data throughput rates (up to 16 MB/sec) if hardware data compression was enabled during recording. The compression feature was enabled during testing, but the degree of data compression realized in each case depended on the specific data used and, in turn, determined the data rate above the manufacturer's rating.

Two SGI multiprocessor Challenge servers, each configured with six CPUs and 512 MB of memory and each running a copy of AMASS, have one PowderHorn silo allocated to each of them. The control of the robotic mechanism of the silo (loading and unloading of the tapes) is via the STK Automated Cartridge System Library Software (ACSLS) running on a SUN SPARC5 workstation. AMASS addresses the ACSLS through a network connection. The ACSLS then controls the robot directly via an RS232 line.



Note: STMGT - ECS Storage Management Code Performing the Archive Control

Figure 2. Archive Hardware and Software Configuration Under Test

All the work described here is done on one of the two silo/host/RAID suites. The final configuration is to be applied to the second suite before proceeding to system integration and test. Therefore, for the remainder of this paper, a single silo, host, RAID unit, etc. will be referred to as the *target configuration.* The target equipment and software configuration is illustrated in Figure 2, GSFC DAAC Archive Repository Equipment Suite.

To date, all of the tuning and configuration adjustments for improving the performance of the archive component were made on either the AMASS software itself, or on the SGI RAID used as a temporary cache, for the data passing to and from the archive silo. Therefore, *AMASS Tuning* is often used as a synonym for the tuning of the overall archive component. STK hardware required no adjustments beyond maintenance.

3 Performance Tuning Methodology and Tools

The stand-alone performance of the archive was assessed by measuring the data throughput to and from the tape drives in the robotic repository. At the time of writing of this paper (November 1997), the data was transferred locally. The only network transfers were of a very small volume of robotic control signals and physical inventory synchronization data between the AMASS software and the STK ACSLS robotic control software.

Figure 3, Data Flow to Tape, shows the test data flow initiated from a directory on a file system disk and directed into the archive. Data Flow from Tape is in the opposite direction. AMASS cache area is a portion of the disk dedicated exclusively for use by AMASS for the maintenance of open archive files and file staging for writes and reads to the archive [1].

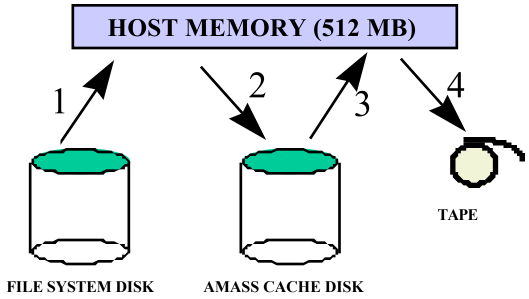
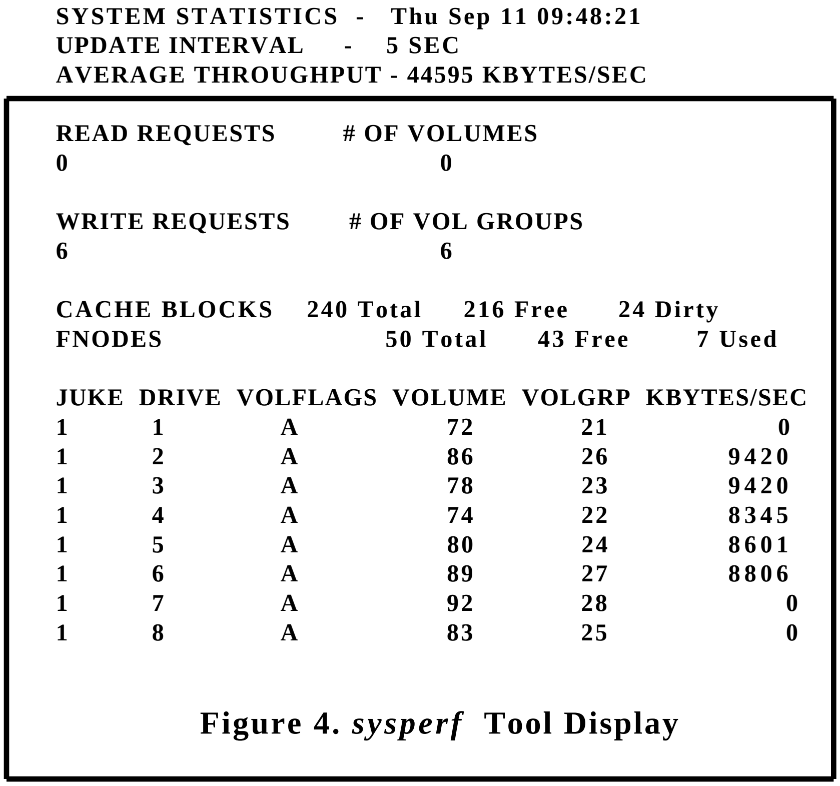


Figure 3. Data Flow To Tape

The most useful tool in assessing throughput performance was the *sysperf* tool supplied by EMASS as part of the software bundle. The tool produces the display shown in Figure 4.

Visibility into AMASS cache area is beneficial for the performance improvement exercises, as well as for problem resolution. Section 5, Software adjustments, describes how cache configuration can be tuned.



```
SYSTEM STATISTICS  -   Thu Sep 11 09:48:21
UPDATE INTERVAL     -    5 SEC
AVERAGE THROUGHPUT - 44595 KBYTES/SEC

READ REQUESTS        # OF VOLUMES
0                                0

WRITE REQUESTS       # OF VOL GROUPS
6                                6

CACHE BLOCKS    240 Total     216 Free     24 Dirty
FNODES                    50 Total      43 Free       7 Used

JUKE  DRIVE  VOLFLAGS  VOLUME  VOLGRP  KBYTES/SEC
1       1          A           72         21           0
1       2          A           86         26        9420
1       3          A           78         23        9420
1       4          A           74         22        8345
1       5          A           80         24        8601
1       6          A           89         27        8806
1       7          A           92         28           0
1       8          A           83         25           0
```

Figure 4. *sysperf* Tool Display

As can be seen from Figure 4, both the individual throughput of a tape drive (under the heading *KBYTES/SEC*) and the overall throughput of the server to the silo can be assessed for the sampling interval. The measurements are taken by EMASS software and reflect its use and freeing of its raw cache segments, rather than actual measurements at the tape drive. The sampling interval can be adjusted in increments of a second from 5 to 60 seconds.

Another very valuable tool was the SGI *Performance Copilot* (*pcp*). The tool afforded an understanding of the RAID disk functioning and facilitated proper reconfiguration of the RAID to maximize throughput.

Finally, UNIX *timex* command was used extensively to gauge the efficiency of a write to or read from AMASS. The data produced by *timex* for a single write to the archive indicated the rate at which user data was being transferred out of the file system disk or into it. *sysperf* timed the outgoing write data streams from AMASS cache on RAID to the tape drives. When multiple background requests were initiated, *timex* could be relied upon to measure the average performance of the transfers to disk and to tape, rather than to memory. That was due to the cumulative size of the transfer exceeding the 512 MB of system memory.

In order to have a measure of control over the read and write data streams, volume *groups* (shown in Figure 4 as *VOLGRP*) were used extensively. A small number of tapes and a separate UNIX directory were assigned to each of the volume groups. This kind of partitioning of the storage allowed data stream direction to one or more tapes, as desired. It was useful for stress testing both the hardware and the FSMS. Simple *Perl 5* scripts with loops issuing sequential *dd* commands were used for initiating the data transfers to and from the archive.

The following sections summarize the performance improvements and give a high level description of the configuration changes that produced them.

4 Initial and Current Performance Comparison

Table 1, Individual Tape Drive Throughput Improvements, and Table 2, Cumulative Archive Throughput Improvements, serve to illustrate data rate performance changes as the result of the integration/tuning activity.

Individual Rate	January 1997 *	August 1997 *
Peak Write (MB/sec)	~ 2	16
Peak Read (MB/sec)	~ 2	16

*Note: Compression is enabled.

Table 1. Individual Tape Drive Throughput Improvements

As mentioned previously, data compression was enabled on the tape drives. The peak performance figure in Table 3 is a reflection of data compression.

Cumulative Rate	January 1997 *	August 1997 *
Peak Write (MB/sec)	7	47
Peak Read (MB/sec)	9.5	29

*Note: Compression is enabled.

Table 2. Cumulative Archive Throughput Improvements

The rise in the peak performance rates is dramatic. It must be noted, that for larger files or for groups of files read from or written to the same tape volume such individual peak streaming performance may be sustained for longer periods of time than for shorter files that are found on a tape or directed to a tape one at a time. Again, as can be expected, individual throughput drops to the cumulative performance limit for multiple concurrent streams. Cumulative performance is very much a function of total concurrent accesses of the AMASS cache disk partition. The peak cumulative throughput to and from the tape drives is measured at times when no *write* transfers from the system disk to the AMASS cache are taking place. Depending on the overall load and where the individual drives are in the load and search cycle, the rates at any instant of time may vary anywhere between zero and the peak. The average, while a function of obtainable peak rates, will also depend on the system use profile.

5 Summary of the Tuning and Configuration Adjustments Efforts

5.1 RAID Adjustments

Summarized below are adjustments that led to the improved peak rates. The greatest measure of performance improvement, due to the enlarging and tuning of the RAID configuration, was realized in the cumulative performance of the system. It must be noted that during the design phase, a RAID configuration rather than *UNIX striped* disk was selected to maximize the reliability of the data. With the projected data rates, and given such large numbers of disk, that would raise the failure incidence, and every failure of non-RAID disk would impose further load on the system. Disk striping was essential in order to meet the data rates. The top throughput rate of the disk configuration is one of the two factors that determine the ceiling on the peak cumulative throughput of the stand-alone archive configuration.

Aside from the sheer data flow rates, another factor affecting throughput performance is the degree of utilization of the tape drives in the robotic silo for reading and writing of the data, as opposed to the drives being idle during tape fetches or spending time for tape mounts, dismounts, rewinds and positioning to the beginning of data. This factor is, to a large extent, an immutable function of the selected tape drive and robotic equipment. An improvement may be realized by trying to maximize data co-location on the tapes or, conversely, data dispersal in the archive. This portion of the work, involving analysis of the processing profiles, is not in the scope of the current paper. It is the former factor, the throughput of the RAID that was targeted by the tuning process.

Figure 5, Expected Data Flow through the FSMS RAID During Operation, is a diagram of the currently anticipated archive data flows on RAID to and from the tape drives once the system comes on-line. As can be seen in the diagram, due to the combination of the AMASS buffering in the *AMASS Cache* and ECS system design requirement to place data in the area labeled as *Staging*, most data is *double hopped* on the RAID. The result is, in the worst case, two writes to and two reads from the same RAID configuration for each archive access.

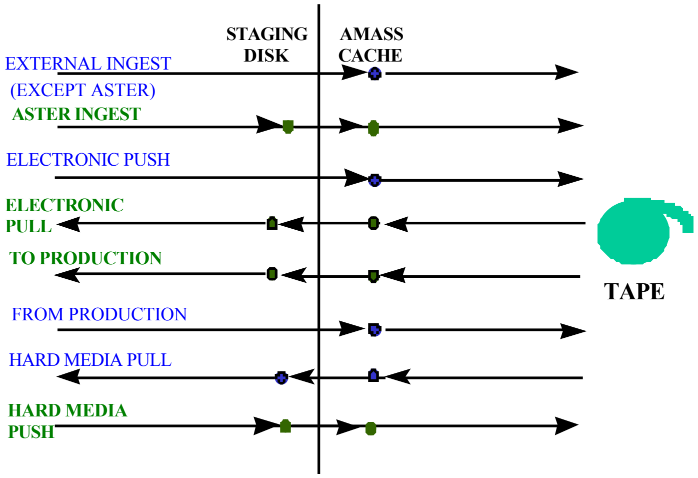


Figure 5. Expected Data Flow through the FSMS RAID during Operation

The change in the RAID configuration level from 5 to 3 accounted for the initial improvement in the individual throughput to and from tape. Most of the individual tape drive performance improvement was due to changes in the AMASS software that allowed for a configurable blocking factor on the tape.

Figure 6, Hardware Configuration under Test as of January 1997, illustrates the hardware configuration that produced the initial throughput rates listed in the January *1997* columns of Tables 1 and 2. As shown, the RAID level configured initially was level 5. Command Tag Queuing (CTQ) was not enabled. Command Tag Queuing, when enabled, provides for request queuing to the disks and request interleaving. The drives were formatted with the default 4 KB block size.

The available RAID disk was divided into 6 GB of raw disk for AMASS Cache with the remainder for the user file system. The stripe element used was 256K blocks. Each block is 512 Bytes. During a test of a write to tape a data file was copied from a directory in the *Staging* user file system partition to the raw *AMASS Cache* partition and subsequently to tape. In the read from tape test the direction of data is reversed (in the direction opposite to the one in Figure 3).

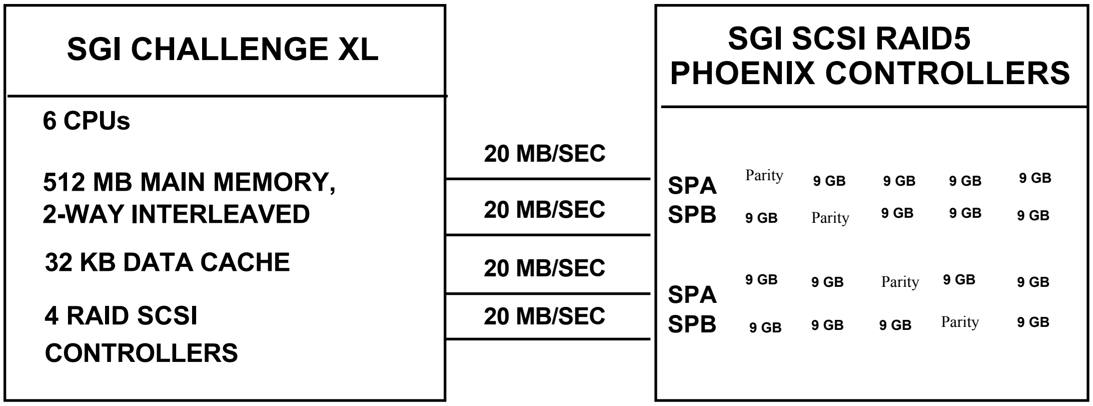

SGI CHALLENGE XL
6 CPUs
512 MB MAIN MEMORY, 2-WAY INTERLEAVED
32 KB DATA CACHE
4 RAID SCSI CONTROLLERS
20 MB/SEC
20 MB/SEC
20 MB/SEC
20 MB/SEC
SGI SCSI RAID5 PHOENIX CONTROLLERS
SPA Parity 9 GB 9 GB 9 GB 9 GB
SPB 9 GB Parity 9 GB 9 GB 9 GB
SPA 9 GB 9 GB Parity 9 GB 9 GB
SPB 9 GB 9 GB 9 GB Parity 9 GB

Figure 6. Archive Hardware Configuration Under Test as of January, 1997

An enlarged RAID configuration currently in use is shown in Figure 7, Archive Hardware Configuration Under Test as of August, 1997. The additional four controllers serve to expand the bandwidth as well as to reduce contention due to multiple disk hits. The same is true for the additional disk capacity.

Aside from physically enlarging the RAID configuration and adding four more SCSI RAID controllers, other RAID configuration parameters were adjusted as follows. 1) Command Tag Queuing (CTQ) was enabled on the RAID with the CTQ depth set to 24; 2) The stripe element for both the raw and the file system portions of the disk was set at 1024 K blocks, thus enlarging the amount of data transferred per disk access; 3) Block size of 64 KB was used to format the drives with *mkfs_xfs*; 4) The raw and the file system portions, were each sized 128 GB and allocated to a set of four separate RAID controllers to minimize disk hit contention; and 5) in sequencing the SCSI address allocation to the A and B controllers or Storage Processors (SPs) during disk partitioning,, the entire A side was listed first, then the entire B side, in order to minimize the detrimental effects of A/B interaction. Thus, when the A side is under load, the B side experiences the resonance effects but is not under load at that time and vice versa.

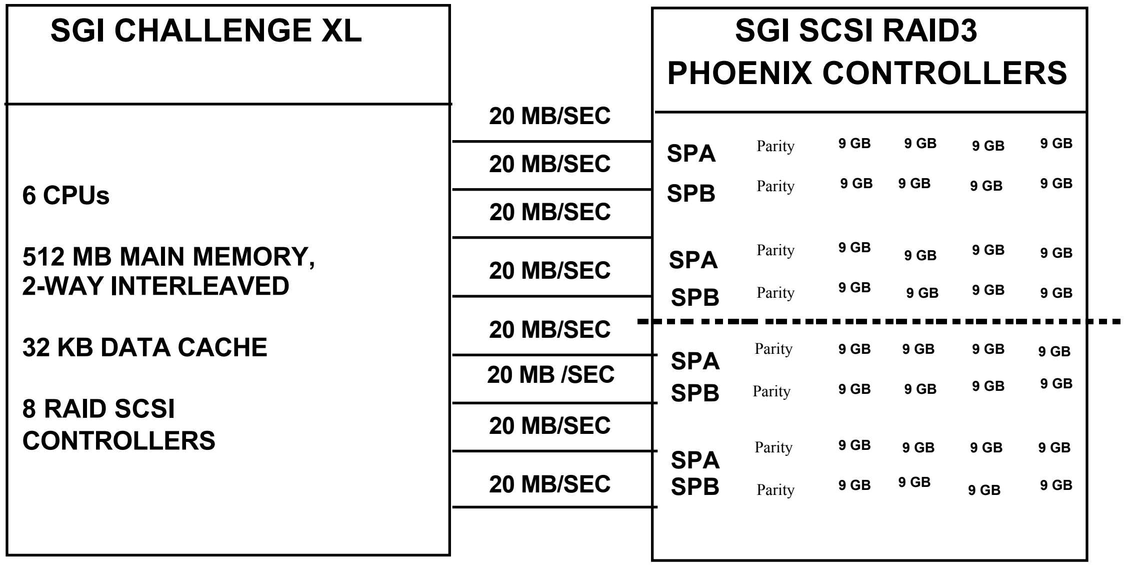

SGI CHALLENGE XL
6 CPUs
512 MB MAIN MEMORY,
2-WAY INTERLEAVED
32 KB DATA CACHE
8 RAID SCSI
CONTROLLERS
20 MB/SEC
20 MB/SEC
20 MB/SEC
20 MB/SEC
20 MB/SEC
20 MB /SEC
20 MB/SEC
20 MB/SEC
SGI SCSI RAID3
PHOENIX CONTROLLERS
SPA Parity 9 GB 9 GB 9 GB 9 GB
SPB Parity 9 GB 9 GB 9 GB 9 GB
SPA Parity 9 GB 9 GB 9 GB 9 GB
SPB Parity 9 GB 9 GB 9 GB 9 GB
SPA Parity 9 GB 9 GB 9 GB 9 GB
SPB Parity 9 GB 9 GB 9 GB 9 GB
SPA Parity 9 GB 9 GB 9 GB 9 GB
SPB Parity 9 GB 9 GB 9 GB 9 GB

Figure 7. Archive Hardware Configuration Under Test as of August, 1997

Table 3, RAID3 Benchmarks, is a summary of *timex* measurements of reads and writes against RAID3 configuration before and after the last tuning adjustments as of November, 1997. Request size of 1024 KB was used.

	April 1997	October 1997
Best Write (MB/sec)	16	50 (file system); 35 (raw)
Best Read (MB/sec)	30	75 (file system); 80 (raw)

Table 3. RAID Benchmarks

5.2 Software Adjustments

5.2.1 Understanding the correct use of AMASS

Some of the initial throughput problems were the result of the lack of user sophistication in the use and configuration of the product. The standard product documentation requires extensive study and discussions with vendor for configuration. The manuals are geared more towards low volume - performance - indifferent sites, versus a facility like an ECS DAAC. However, the advanced user training offered by the vendor was most helpful and technical support of the configuration changes informative. The following adjustments were made in the ECS use of AMASS.

1) Use of *dd* data copy. Surprisingly low initial throughput results can be partially attributed to the use of the UNIX copy command, *cp,* for data transfers to and from AMASS directories. The block size used by the *cp* command for data transfers was initially 4 KB. Such transfers were, understandably, very slow. The *dd* command with a block size of 1024 KB is now used for all AMASS data transfer tests and tuning.

2) Appropriate tuning of AMASS cache. AMASS cache parameters were tuned to the prevalent file size in order to optimize recording of that file to tape. As an example, in order to optimize the transfer of a 1 GB file to tape, each of the four segments comprising the total portion of data going to tape at one time was made to be 256 KB. Such tuning for larger files would result in excessive disk allocation if smaller files prevailed in the system.

This adjustment was required due to the effect that the tape drive buffer flush delay had on the total write to tape throughput. Disk transfer would block for the particular process while its tape drive buffer was being flushed. Since a single buffer flush, lasting for up to 20 seconds, occurred after each four *dirty cache blocks* written to tape, as well as at the end of a file, it was important to minimize the number of such flushes.

Cache size tuning to improve write performance for large files can, however, limit the number of simultaneous processes which can use that cache (the NFNODE parameter, that the vendor defines, as the number of files that could be open in the AMASS file system at the same time [1], is inversely related to the cache block size). Fortunately, in the near future, the vendor is discontinuing the synchronous buffer flush to tape and is going to an asynchronous buffer flush. This will not only greatly improve the performance of an individual write to tape, but also make the fine adjustment to the prevalent file size unnecessary.

5.2.2 Performance Related AMASS corrections by the Vendor

In the interval from the time when the testing of AMASS on the target configuration commenced, at the start of 1997, and the time of this writing, November of 1997, several product deficiencies with respect to data throughput were corrected. Each major correction, as well as resolution of other non-design software problems as encountered, was accompanied by extensive testing on the ECS side, stretching to a year of near-full time involvement by the author and other team members as necessary. The three significant performance related corrections are detailed below:

1) Adjustable size blocking factor to tape. Initially, the size of the data block written to tape was *hardwired* at 16 KB. The drive manufacturer, STK, recommends a block size of 256 KB for maximum throughput. Configurable blocking factors for tape writes contributed to the major performance improvement, as can be seen from Tables 1 and 2.

2) Asynchronous library operation. When first tested, AMASS did not allow for asynchronous STK PowderHorn library operation on mount and dismount; no library activity took place during the time when any one of the drives was performing a tape load, positioning, or rewind. Needless to say, the associated time loss was unacceptable. The correction now allows for a fully asynchronous rewind and unload of tapes.

The asynchronous tape mount correction was accompanied by time-out problems because the hardwired retry parameter did not quite account for the time required to rewind a 50 GB tape cartridge. There is now an adjustable parameter controlling the number of retries of the drive access after the tape mount. Having configurable control over the number of retries solved the time-out problem on the mount side.

3) Introduction of Asynchronous I/O. Asynchronous I/O contributed to raising the overall performance ceiling by allowing multi-threaded transfers to AMASS cache and from it.

4) Removal of a 2 GB limit on the size of a single AMASS cache partition. Beginning with Version 4.9 for IRIX 6.2, AMASS supports a maximum total cache partition size of 1 TB,

and the total cache size up to 2 TB in as few as two partitions[2]. Although the larger cache partition size has no direct effect on the individual channel throughput rate, large total cache size allows larger number of simultaneous archive processes. In turn, the large number of simultaneous reads and writes affords a greater queuing and multithreading efficiency.

6 Conclusion

The paper contains an overview of configuration adjustments and the corresponding performance improvements in the GSFC ECS DAAC archive configuration. The information presented here is a snapshot for November, 1997 and represents work in progress. Further performance improvement are expected in December, 1997, as a result of removal of a 2 GB per disk partition limitation and elimination of tape drive buffer flush time expenditure. Although the tuning described here took place on a specific set of equipment and software, the functional components involved are generic to large archive systems. High capacity/high performance systems are becoming more common, and, increasingly higher performing hardware is becoming available. It is hoped that our experience may be of use to integrators with similar architectures.

7 Acknowledgements

The author wishes to acknowledge Stacey Brown and Brad Koenig of Lockheed Martin Space Mission Systems and Services, and Byron Peters of Hughes Information Systems for their invaluable contributions in getting the archive configuration to the performance it has achieved today. Appreciation is due to Randy Kreiser of SGI, and Andy Richards and Leo Profilet of EMASS for guidance with their respective products. And, finally, the author thanks Robert Howard of Hughes Information Systems, Ray Simanowith of Lockheed Martin Space Mission Systems and Services, and Jean-Jacques Bedet of Hughes STX for the review of the text.

8 References

[1] Installing AMASS, EMASS, Inc. Version 4.8, December 1996, Document Number 600422

[2] Release Notes for AMASS Version 4.9, November 14, 1997, EMASS, Inc. Document Number 600459

[1] ECS is developed by the National Aeronautics and Space Administration (NASA) through Contract NAS5-60000 as part of NASA's Mission to Planet Earth.

[2] Here and subsequently, AMASS and EMASS are either trademarks or registered trademarks of EMASS, Inc. Silicon Graphics, Challenge, IRIX, and Performance Copilot are registered trademarks of Silicon Graphics, Inc. PowderHorn, RedWood (SD-3), and ACSLS are trademarks of Storage Technology Corporation. SPARC5 is a trademark of SUN Microsystems, Inc. UNIX is a trademark of AT&T Bell Laboratories.